BINUS University

Academic Career: ***Undergraduate / ~~Master / Doctoral / BINUS Online/ Professional~~*)***	**Class Program:** ***Regular/ ~~Global~~ Class*)***
☒ **Mid Exam** ☐ **Others Exam : _____** ☐ **Final Exam**	**Term : Odd / ~~Even / Compact~~*)** **Period (Only for *BINUS Online/ Master*): 1 / 2*)**
☒ **Kemanggisan** ☐ **Senayan** ☐ **Semarang** ☐ **Alam Sutera** ☐ **Bandung** ☐ **Medan** ☐ **Bekasi** ☐ **Malang** ☐ **BiOn**	**Academic Year :** **2025 / 2026**

Exam Type*	:	~~Onsite~~ / ~~Online /~~ Take Home	Faculty / Dept.	:	SoCS / Data Science
Day / Date**	:	Thursday/13 Nov 2025	Code - Course	:	DTSC6008001- Text Mining
Time**	:	13:00			
Exam Specification***	:	☐ Open Book ☐ Open Notes ☐ Close Book ☐ Oral Test ☐ Open E-Book	Class	:	Regular
Equipment***	:		Student ID ***	:	
☐ Examination Booklet ☐ Calculator ☐ Dictionary		☐ Laptop ☐ Drawing Paper – A3 ☐ Tablet ☐ Drawing Paper – A2 ☐ Smartphone ☐ Notes: _____sheet	Name ***	:	
			Signature ***	:	

*) *Strikethrough the unnecessary items* **) *For Online Exam, this is the due date* ***) *Only for Onsite Exam*

Please insert the test paper into the examination booklet and submit both papers after the test.*

The penalty for CHEATING is DROP OUT!

Learning Outcome for

LO1: Identify the scope of text mining problem
LO2: explain the fundamental text mining theory
LO3: apply text mining techniques with Python
LO4: analyze the results of text mining process

Panduan

1. Tuliskan jawaban anda di python notebook, dengan ketentuan sebagai berikut:
   a. Gunakan code cell untuk menerapkan algoritma atau kalkulasi.
   b. Gunakan Markdown cell untuk menuliskan penjelasan
   c. Jika diperlukan, anda dapat menyisipkan gambar dengan menggunakan fitur 'insert image' yang tersedia di Menu Edit.
   d. Anda bisa menggunakan Machine Learning atau Text Library seperti SKLearn, Pytorch, Tensorflow-Keras, NLTK dan Gensim.
2. Berikan penjelasan untuk setiap jawaban anda dalam bentuk video (lampirkan **link video** pada file .ipynb). Letakkan tautan video di python notebook. Harap dipastikan tautan dapat diakses publik.
3. Berkas yang dikirimkan dalam bentuk ZIP (NIM.zip) yang terdiri atas 2 berkas, yaitu:
   a. file data .csv
   b. file .ipynb

Verified by Department,

[Noviyanti TM Sagala] (D6464)
10 21, 2025

Studi Kasus

Untuk mengisi waktu luang, anda suka mencoba berbagai games yang tersedia di Playstore. Suatu ketika sebelum mendownload suatu games, anda membaca review-review yang terdapat di Playstore mengenai games tersebut beserta ratingnya. Anda berpikir jika dapat membangun model prediksi rating, model ini dapat digunakan pihak games untuk memprediksi konten dari sumber lain. Maka anda berinisiatif untuk:

1. **[LO 1, LO 2, LO 3 – 10 Points]** Mengumpulkan data review games dengan tema yang anda suka dengan cara scrapping sebanyak 1000 data dengan berbagai nilai rating. Anda juga memutuskan untuk mengumpulkan data review dalam bahasa Inggris.

2. **[LO 1, LO 2, LO 3 – 10 Points]** Anda ingin mengenal data yang anda scrapping, maka anda melakukan exploratory terhadap data untuk melihat kata-kata yang dominan disetiap rating dan memeriksa apakah kata-kata yang tidak sesuai dengan standar bahasa punya frekuensi yang besar

3. **[LO 1, LO 2, LO 3 – 15 Points]** Anda melihat bahwa data ini tidak bisa langsung digunakan dalam membangun model prediksi rating, maka anda melakukan preprocessing data hingga mendapatkan bentuk token paling efektif yang sesuai kaidah tata bahasa dan sebisa mungkin makna kata tersebut tidak berubah apalagi kelas katanya.

4. **[LO 1, LO 2, LO 3 – 10 Points]** Anda memutuskan akan menggunakan machine learning unutk model prediksi, namun sebelum itu anda perlu merepresentasikan data teks ini kedalam bentuk vektor yang dapat diterima oleh machine learning sebagai input, maka anda menerapkan 2 bentuk metode text representation yaitu, metode yang nilai vectornya merepresentasikan seberapa penting kata tersebut dalam suatu data sample, dan metode yang vectornya direpresentasikan oleh kata disekitarnya yang ditraining dengan model ANN dengan input kata dan target kata yang ada disekitarnya. Anda akan mencoba kedua metode text representation ini sebagai input yang kemudian akan dibandingkan performance prediksinya.

5. **[LO 1, LO 2, LO 3 – 15 Points]** Setelah text represntation selesai, anda melakukan pemodelan prediksi dengan menggunakan 2 metode Machine Lerning yang anda pilih sendiri; tentu saja anda juga perlu melakukan tuning hyperparameter minimal 2 hyperparameter untuk masing-masing algoritma machine learning.

6. **[LO 1, LO 2, LO 3, LO4 – 10 Points]** Untuk mengetahui performa model prediksi yang anda kerjakan, tentu anda perlu melakukan performance evaluation dan juga menganalisa hasilnya. Anda melakukan perbandingan performance test data dari model yang dibuat dengan metode text representation yang berbeda, dan anda membuat summary hasil sebagai berikut:

Text Representation	Algoritma Machine learning	Machine learning Hyperparameter	Metric Evaluation 1	Metric Evaluation 2	Metric Evaluation 3	Metric Evaluation 4
Metode 1	ML 1	xxx xxx	xxx xxx	xxx xxx	xxx xxx	xxx xxx
Metode 1	ML2					
Metode 2	ML1					
Metode 2	ML2					

Verified by Department,

[Noviyanti TM Sagala] (D6464)
10 21, 2025

7. **[LO 1, LO 2, LO 3, LO 4 – 20 Points]** Anda mendapati ternyata data anda imbalance, sehingga anda penasaran, apakah jika kondisi imbalance data ini dihandle, akan ada perubahan pada perfomance atau tidak, sehingga anda melakukan treatment terhadap data imbalance dan melakukan training kembali dan melakukan evaluasi terhadap model dengan kondisi data baru. Anda membandingkan evaluasi performancenya dan mendapatkan kesimpulan, model mana yang terbaik.

8. **[LO 1, LO 2, LO 3, LO 4 – 10 Points]** Untuk setiap proses di atas anda memberikan penjelasan dengan detail baik pada dokumen Notebook maupun pada Video agar dapat bermanfaat bagi orang lain.

Verified by Department,

[Noviyanti TM Sagala] (D6464)
10 21, 2025